Decoupling Vocal Tract from Glottal Source Estimates in Speaker's Identification

Pedro Gómez, Agustín Álvarez, Luis Miguel Mazaira, Roberto Fernández, Víctor Nieto, Rafael Martínez, Cristina Muñoz, Victoria Rodellar

Grupo de Informática Aplicada al Procesado de Señal e Imagen, Facultad de Informática, Universidad Politécnica de Madrid, 28600 Boadilla del Monte, Madrid, e-mail: pedro@pino.datsi.fi.upm.es

Abstract

Classical parameterization techniques in Speaker Identification tasks use the codification of the power spectral density of speech as a whole, not discriminating between articulatory features due to the dynamics of vocal tract (acoustic-phonetics) and those contributed by the glottal source. Through the present paper a study is conducted to separate voicing fragments of speech into vocal and glottal components, dominated respectively by the vocal tract transfer function estimated adaptively to track the acoustic-phonetic sequence of the message, and by the glottal characteristics of the speaker and the phonation gesture. In this way information which is conveyed in both components depending in different degree on message and biometry is estimated and treated differently to be fused at the time of template composition. The methodology to separate both components is based on the decorrelation hypothesis between vocal and glottal information and it is carried out using Joint Process Estimation. This methodology is briefly discussed and its application on vowel-like speech is presented as an example to observe the resulting estimates both in the time as in the frequency domain. The parameterization methodology to produce representative templates of the glottal and vocal components is also described. Speaker Identification experiments conducted on a wide database of 240 speakers is also given with comparative scorings obtained using different parameterization strategies. The results confirm the better performance of de-coupled parameterization techniques compared against approaches based on full speech parameterization.

Introduction

Forensic Sciences in relation with Security Applications are experiencing a strong expansion nowadays. Among the different technologies included into the term, Speech Processing and Recognition is one of most relevance due to the inlcreasing role and ubiquity that these play in Human Communication.

The widespread use of new technologies (cellular telephony, IP voice, voice-managed services, etc.) make the task even more complex. The present paper is oriented to explore the applicability of basic Speech Processing capabilities to the Forensic Analysis of Voice [8] as reflected in Figure 1.

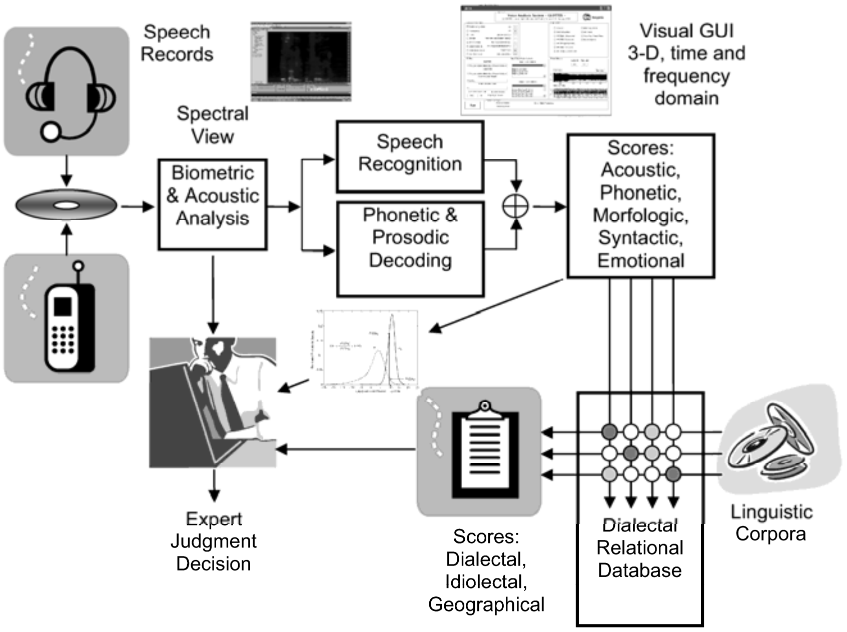
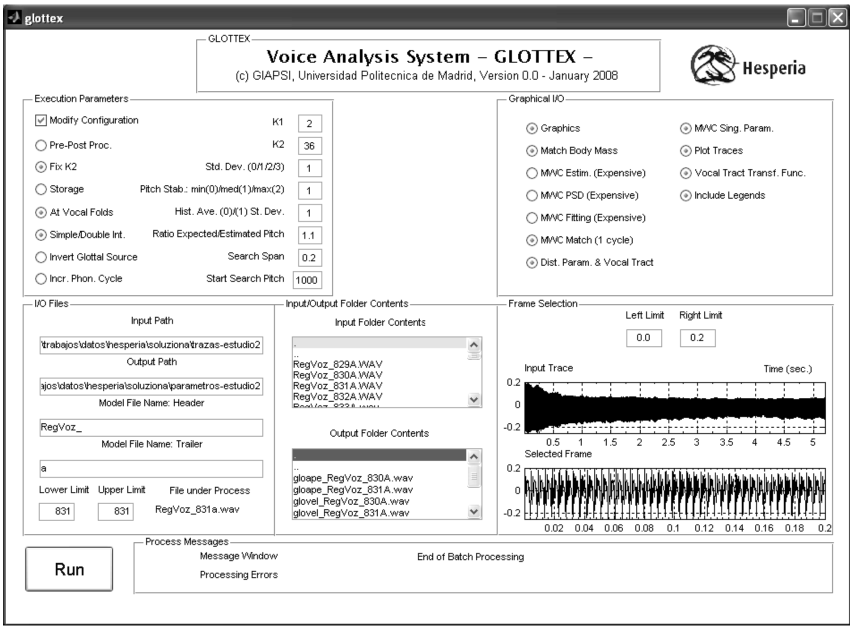

Speech Records
Spectral View
Visual GUI 3-D, time and frequency domain
Biometric & Acoustic Analysis
Speech Recognition
Phonetic & Prosodic Decoding
Scores: Acoustic, Phonetic, Morfologic, Syntactic, Emotional
Expert Judgment Decision
Scores: Dialectal, Idiolectal, Geographical
Dialectal Relational Database
Linguistic Corpora

Figure 1. Experimental framework in Speech Processing for Forensic Studies

The main premise in any Forensic Speech Analysis framework is ultimately oriented to facilitate expert judgment. This is due to the enormous complexity of the process, involving analysis from the integrity of the recordings and media to high levels of linguistic and semantic analysis. The judgment demands large amounts of information which is to be supplied to an expert or a team of experts with enough accreditation in the capabilities required to produce reliable reports. Under this point of view the role of any Experimental Framework would be that of providing the expert with enough information at the different levels where judgment is to be supported upon: Acoustic-Phonetic, Phonologic, Morphologic, Prosodic, Emotional, Idiolectal, Dialectal, etc. These rich levels will demand the strong co-operation of Engineering and Linguistics in close contact, therefore multidisciplinarity is mandatory. These considerations are reflected in the sketch of Figure 1, where speech recorded or stored is deconstructed under the first levels of analysis: Biometrical and Acoustical, in such a way that the information extracted from these levels can be directly used by experts, as well as fed to other classification levels. These may include Speech, Speaker and Emotion Detection, Phonetic-Acoustic Decoding, Broad Class Phonetic Labelling, etc. These tasks may be fulfilled by more or less automatic pattern recognition engines. The information supplied at these different levels will be scored and supplied to the expert as LR statistics or similar for further judgment. Other aspects of the analysis will include automatic classification under the Idiolectal, Dialectal and Geographical Information Systems using Data Mining Techniques. These scores will also be supplied for further judgment.

Separating Glottal Source and Vocal Tract

The present work will summarize briefly the possibilities offered at the Biometrical and Acoustical levels by the application of voiced speech analysis tools capable of splitting vocal from glottal information [3]. The Graphical User interface of one such tool is shown in Figure 2.



glottex

GLOTTEX

Voice Analysis System – GLOTTEX –

(c) GIAPSI, Universidad Politecnica de Madrid, Version 0.0 - January 2008

Hesperia

Execution Parameters

☑ Modify Configuration — K1: 2
○ Pre-Post Proc. — K2: 36
● Fix K2 — Std. Dev. (0/1/2/3): 1
○ Storage — Pitch Stab.: min(0)/med(1)/max(2): 1
● At Vocal Folds — Hist. Ave. (0)/(1) St. Dev.: 1
● Simple/Double Int. — Ratio Expected/Estimated Pitch: 1.1
○ Invert Glottal Source — Search Span: 0.2
○ Incr. Phon. Cycle — Start Search Pitch: 1000

Graphical I/O

● Graphics
● Match Body Mass
○ MWC Estim. (Expensive)
○ MWC PSD (Expensive)
○ MWC Fitting (Expensive)
● MWC Match (1 cycle)
● Dist. Param. & Vocal Tract
● MWC Sing. Param.
● Plot Traces
● Vocal Tract Transf. Func.
● Include Legends

I/O Files

Input Path: \trabajos\datos\hesperia\soluziona\trazas-estudio2
Output Path: ajos\datos\hesperia\soluziona\parametros-estudio2
Model File Name: Header: RegVoz_
Model File Name: Trailer: a
Lower Limit: 831
Upper Limit: 831
File under Process: RegVoz_831a.wav

Input/Output Folder Contents

Input Folder Contents
```
.
..
RegVoz_829A.WAV
RegVoz_830A.WAV
RegVoz_831A.WAV
RegVoz_832A.WAV
```

Output Folder Contents
```
.
..
gloape_RegVoz_830A.wav
gloape_RegVoz_831A.wav
glovel_RegVoz_830A.wav
glovel_RegVoz_831A.wav
```

Frame Selection

Left Limit: 0.0
Right Limit: 0.2

Run

Process Messages

Message Window: End of Batch Processing
Processing Errors

Figure 2. Graphical User Interface of the tool GLOTTEX® glottal source separation and parameterization [7]

The advantages of this split open the possibility of independently studying vocal and glottal components. It is well known that the vocal tract transfer function expressed by its resonances (formants) is of great interest for the biometrical characterization of the speaker [5][8][9]. The glottal source descriptions in the time or frequency domain are well known for their capability of expressing speech pathology [2]. But as both correlates, vocal and glottal, appear intermingled in the acoustic recording of speech, techniques relying on the analysis of the acoustic record of full voice resent from this juxtaposition and blurring and become less efficient. A good approach will be to split voice into vocal tract and glottal source information for further analysis with current automatic pattern recognition engines.

Materials and Methods

The splitting of voiced speech into glottal and vocal components is a fruitful methodology, as a brief examination to Figure 3 will help us to conclude. The template presents the formants of a voiced speech segment extracted from Linear Predictive Coding as used in vocal from glottal separation. On top (dot line), the power spectral density of the acoustic record (full voice) may be observed. It is relatively common to see this plot mistaken as the vocal tract transfer function, but this is not true anymore. The first peak observed and encircled corresponds to the so called "glottal formant", and thus can not be considered contributed by a true vocal tract resonance. When the glottal source is removed the bottom curve (thin line) can be associated to the vocal tract transfer function. Interestingly, the first true formant of the vocal tract (also encircled) is now lower in strength than the second true formant, as the spectral tilt introduced by the glottal source faded this difference. The same may be said from formants at higher frequencies, as for example the one around 10 kHz, which may exhibit an intensity comparable to the lower ones. Therefore formant estimations using this technique will be much more reliable and robust. But these are not the only benefits achievable, as the glottal source analysis is quite rich for biometrical speaker recognition and pathology detection. In fact a work by Plumpe et al. [6]

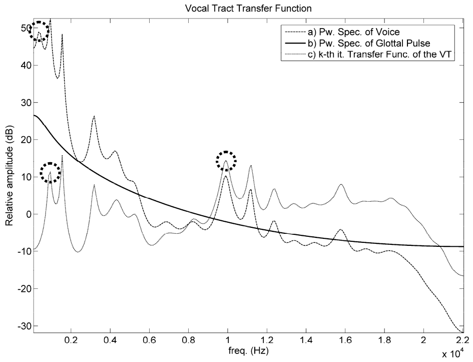


Figure 3. Separation of the vocal tract and glottal source by Adaptive Linear Prediction. It may be seen that the glottal formant has been removed, and that the formants of the input speech trace have been balanced, which higher formants above lower ones.

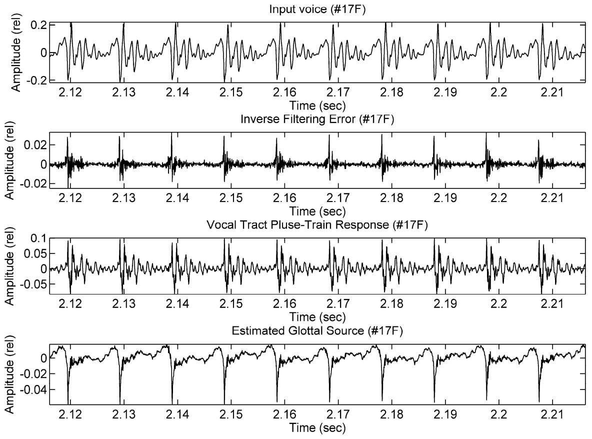


Figure 4. Time domain traces. From top to bottom: input speech $s_l(n)$, inverse filtering error $e_g(n)$, pulse-train response of the vocal tract $s_v(n)$ and glottal estimate $s_g(n)$

signalled this possibility several years ago proposing the extraction of the L-F pattern of the glottal source [1] for speaker identification purposes. The main difficulty found in doing so comes from the production of good glottal source estimates. Voice splitting can grant this, as shown in Figure 4. In this picture the results of applying voice splitting to a phonation trace can be appreciated from top to bottom templates, showing the acoustic voice record, the inverse filtering error residual, the vocal tract impulse response, and the reconstructed glottal source. A close-up view of two cycles of the glottal source for a different phonation are given in Figure 5 for detailed inspection.

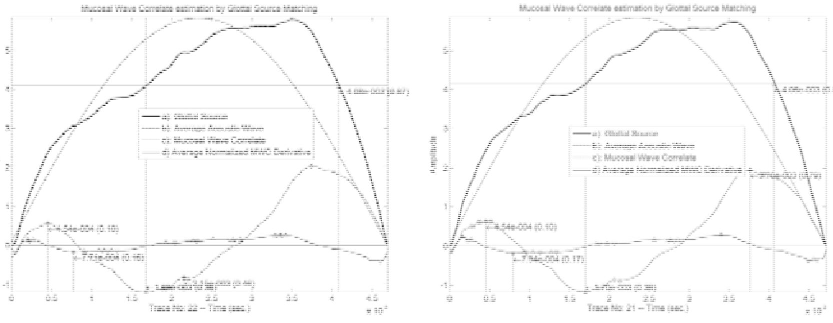


Figure 5. Glottal Source reconstruction (two consecutive cycles), showing the estimates of the Average Acoustic Wave, the Mucosal Wave Correlate and the first derivative of this last signal. The Open and Close Quotients are given as absolute and relative values.

In this case two neighbour glottal source cycles (starting at the glottal closure instant) obtained from GLOTTEX® are plotted for the purpose of

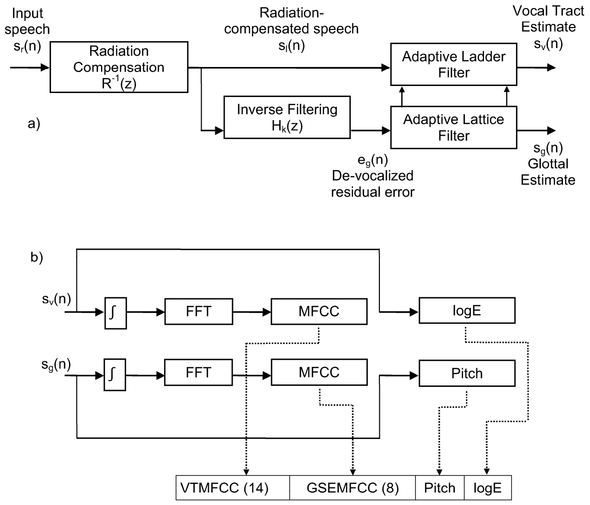


Figure 6. a) General framework to separate vocal from glottal characteristics by adaptive joint estimation. b) Parameterization scheme used in the experiments.

biometrical time-domain analysis. Open and Close Quotients are given as absolute time evaluations or relative to the phonation cycle duration. The Opening and Closing instants are highlighted as well. Aspects as vocal fold asymmetry in vibration may be easily examined. Besides for time-domain analysis GLOTTEX® is designed to offer frequency domain parameterization of the glottal source (glottal signature), common distortion parameters (*pitch, jitter, shimmer, harmonics-noise ratio*), and biomechanical parameters (*fold mass and tension estimates*). The before mentioned separation technique has been also used for speaker identification purposes according to the general framework summarized in Figure 6. In the upper template (a) the signal processing line for vocal and glottal separation is explained. It consists in a radiation compensation filter, followed by an inverse filter to eliminate the vocal tract resonances. The de-vocalized speech is jointly estimated in contrast to the radiation-compensated speech by a lattice-ladder adaptive structure, producing estimates of the vocal tract impulse response and the glottal residual. In the lower template (b) both estimates are integrated and transformed to the frequency domain by FFT. The mel-cepstrum coefficients for both components are obtained, 14 for the vocal component and 8 for the glottal one. Estimates of the pitch and the logarithm of the energy are produced as well. With these parameters a characteristic vector is constructed and tested by a Gaussian Mixture Model Set.

Results and Discussion

The performance of this methodology was checked in the following context: a database including a wide representation of the phonetic articulation and the glottal characteristics for 240 speakers [4] with a good description of intra- and inter-speaker variability. The data set was divided into 176 speakers used for modelling during the training phase, and 64 speakers serving as impostors during the test phase. The training dataset is composed by 10 sentences per each of the modelled speakers comprising approximately 30 sec. of speech. The testing set consisted in 10 sentences per known speaker as well as 10 sentences from each impostor speaker, each sentence lasting 4 sec for both groups. Training and testing sets for each speaker are based on different sentences. The training session produced Gaussian Mixture Models for each modelled speaker of order *k={16, 32, 64, 128, 256, 512}* the testing set was processed in a closed-set setup and the scores recorded in relation to the log-likelihood threshold θ used for each experiment. The baseline procedure for comparison used a characteristic vector composed of 14 mel-cepstrum coefficients plus estimations of pitch and the log of energy and the optimum classifier among the mentioned values of *k* corresponding to order *k=64* was used as contrast. The comparison results are given in Figure 7. It may be seen that GMM model orders of *k=64, 32* or *128* combining vocal and glottal information surpass the best results obtained using full voice. The best equal error rate is of 0.35% False Acceptances and False Rejections.

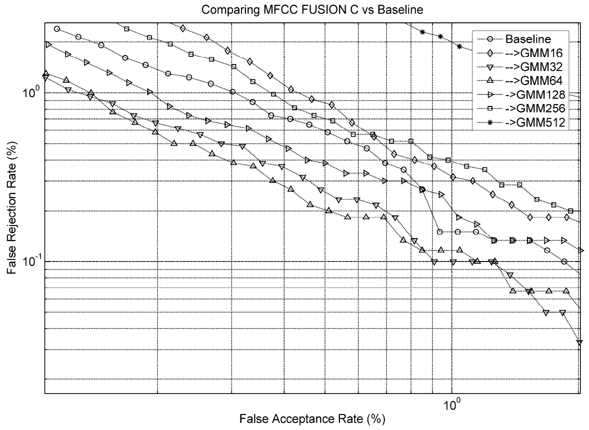


Figure 7. Comparison of the identification results from the combined parameterization of raw speech, vocal tract and glottal source against the baseline (full voice).

Conclusions

Through the present paper the separation of vocal and glottal information present in voiced speech is proposed as a technique to: enhance vocal tract transfer functions based on formant spectra, produce robust estimates of the glottal source and enhance speaker identification results from fusion parameter cocktails including split vocal and glottal estimations. In speaker identification tasks this methodology improves the best method based on raw speech in a 50% over equal error rates. The methodology can be used in biometric systems, forensics and secure access applications, among others.

Acknowledgments

This work is being funded by grants TIC2003-08756 and TEC2006-12887-C02-01/02 from Plan Nacional de I+D+i, Ministry of Education and Science, by grant CCG06-UPM/TIC-0028 from CAM/UPM, and by project HESPERIA (http.//www.proyecto-hesperia.org) from the Programme CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

References

[1] Fant G., Liljentcrants J., Lin Q., “A four-parameter model of glottal flow”, *STL-QSPR*, Vol. 4, 1985, pp 1-13.

[2] Godino, J. I., Gomez, P., “Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors”, *IEEE Trans Biomed. Eng*. Vol. 51, 2004, pp. 380-384.

[3] Gómez, P., et al., “Biometrical Speaker Description from Vocal Cord Parameterization”, Proc. of ICASSP’06, Toulouse, France, 2006, pp. 1036-1039.

[4] Moreno, A., et al., “ALBAYZIN Speech Database: Design of the Phonetic Corpus,” *Proc. Eurospeech’93*, vol. 1, 1993, pp. 175–178.

[5] Nickel, R. M., “Automatic Speech Character Identification”, IEEE Circuits and Systems Magazine, Vol. 6, No. 4, 2006, pp. 8-29.

[6] Plumpe, M. D., Quatieri, T. F., Reynolds, D. A., “Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification”, IEEE Trans. on Speech and Audio Proc., Vol. 7, No. 5, 1999, pp. 569-586.

[7] Project MAPACI: http://www.mapaci.

[8] Rose, P., *Forensic Speaker Identification*, Taylor & Francis, New York, 2002.

[9] Whiteside, S. P., “Sex-specific fundamental and formant frequency patterns in a cross-sectional study,” *J. Acoust. Soc. Am.*, Vol. 110, No. 1, pp. 464–478, 2001.